Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)

Gunarathne W.H.S.D
Department of Software Engineering
Sri Lanka Institute of Information Technology (SLIIT)
Malabe, Sri Lanka
it14029714@my.sliit.lk

Perera K.D.M
Department of Software Engineering
Sri Lanka Institute of Information Technology (SLIIT)
Malabe, Sri Lanka
dulani.p@sliit.lk

Kahandawaarachchi K.A.D.C.P
Department of Information Systems Engineering
Sri Lanka Institute of Information Technology (SLIIT)
Malabe, Sri Lanka
chathurangika.k@sliit.lk

Abstract—**Chronic Kidney Disease (CKD) is considered as kidney damage which lasts longer than three months. In Sri Lanka, CKD has become a severe problem in the present days due to CKD of unknown aetiology (CKDu) that can be seen popularly in North Central Province. Identifying CKD in the initial stage is important to provide necessary treatments to prevent or cure the disease. In this work main focus is on predicting the patient's status of CKD or non CKD. To predict the value in machine learning classification algorithms have been used. Classification models have been built with different classification algorithms will predict the CKD and non CKD status of the patient. These models have applied on recently collected CKD dataset downloaded from the UCI repository with 400 data records and 25 attributes. Results of different models are compared. From the comparison it has been observed that the model with Multiclass Decision forest algorithm performed best with an accuracy of 99.1% for the reduced dataset with the 14 attributes.**

Keywords-Chronic Kidney Disease; symptoms; predictive models; machine learning; classification algorithms

I. INTRODUCTION

Chronic Kidney Disease (CKD) is one of the leading health issues that abound in present day of Sri Lanka. Chronic Kidney Disease of unknown etiology (CKDu) has been rapidly developing in North-Central province agricultural zones in Sri Lanka and by now it has become a severe problem.

The present work emphasize is based on data mining, and the classification techniques in health informatics to detect Chronic Kidney Disease (CKD). Health informatics applies informatics concepts, theories and practices to real-life situations to achieve better health outcomes [1]. **Data Mining** is a cross-disciplinary field that focuses on discovering properties of data sets. **Machine Learning** is a sub-field of data science that focuses on designing algorithms that can learn from and make predictions on the data [2].

Kidney disease occurs when kidneys are unable to perform their functions properly. Kidneys are no longer able to remove waste products and extra water from the blood, which ends up building up in the body, causing several complications. Kidney damage and decreased function that lasts longer than three months is called chronic kidney disease (CKD) [3]. CKD is also defined as either kidney structural damage or a decrease in Glomerular Filtration Rate (GFR) < 60 ml/min/1.73 m2 for three or more months. The common meaning of the above terms is slow progressive loss of kidney functions over time due to progressive destruction of renal mass. Because of this slow progressive loss of kidney function, CKD often goes undetected and undiagnosed until it gets worse slowly over time [3]. If a kidney disease is found and treated early, it will help to slow down the CKD process of going to later stages from the current stage. To recognize the disease early, having a proper knowledge about disease symptoms is very important.

In this research we developed an automated machine leaning solution to detect CKD by exploring 14 attributes related to kidney disease. The dataset used for evaluation consists of 400 individuals and suffers from noisy and missing data. We use these data to build different classification models with different classification algorithms. By using these models we will predict the CKD and non CKD status of the patients. Accuracy results of different models are compared and from the comparison the best model with the relevant algorithm and better accuracy for the given dataset will be selected as the best model to predict the CKD status of patients. This will help to achieve fast accurate results on predictions using machine learning techniques which will lead to reduce the time waste on CKD identification. Early identification of the disease will provide benefits for both doctors and patients in providing early treatments and reducing the speed of disease progression.

TABLE I. STAGES AND ACTION PLAN FOR CKD [4]

2471-7819/17/31.00 ©2017 IEEE
DOI 10.1109/BIBE.2017.00056

Stage	Description	GFR (mL/min/1.73 m2)	Action
-	At increased risk for CKD	>=90 with risk factors	Screening CKD risk reduction
1	Kidney damage with normal or increased GFR	>=90	Diagnosis and treatment Slow progression of CKD Treat comorbidities Cardiovascular disease reduction of risk
2	Mild decrease in GFR	60-89	Estimate progression
3	Moderate decrease in GFR	30-59	Evaluate and treat complications
4	Severe decrease in GFR	15-29	Prepare for renal replacement therapy
5	Kidney failure	< 15 or dialysis	Replacement if uremic

II. LITERATURE SURVEY

Many number of researchers have used Machine Learning and data mining-based algorithms to solve problems in health sector. They have used different techniques and methods to classify and predict the CKD status of the patients.

Charleonnan et al. [5] have evaluated four machine learning methods including K-nearest neighbors (KNN), support vector machine (SVM), logistic regression (LR), and decision tree classifiers to predict CKD. These models are constructed from a CKD dataset which was collected from Apollo Hospitals Indians, and the performance of these models are compared together in order to select the best classifier to predict the chronic kidney disease. From the experimental results they obtained, it shows that SVM classifier has obtained the highest accuracy of 98.3%. In addition to that, SVM has the highest sensitivity after training the dataset.

S. Dilli Arasu and Dr. R. Thirumalaiselvi [6] have carried a research to handle the missing values in a CKD dataset. Having missing values in a dataset will reduce the accuracy of the prediction result. To avoid the missing values they have performed a recalculation process on CKD stages and the unknown values are filled by using the recalculated values. In this research they have proposed Weighted Average Ensemble Learning Imputation (WAELI) algorithm to predict the missing values using both single value and multiple value imputations. The final value is predicted by computing the weighted average of each model. According to the results they observed it proves that that the proposed WAELI and WAELI-FPA techniques are better than the existing algorithm in the terms of accuracy, precision, recall and F-measure.

Sahil Sharma, Vinod Sharma, and Atul Sharma [7], have evaluated 12 classification techniques by applying them to a dataset with 400 instances and 24 attributes. To calculate the efficiency the prediction results of the candidate methods has been compared with the actual medical results of the subject. They have used predictive accuracy, sensitivity, precision and specificity as the performance evaluation metrics. According to the results observed, the decision tree technique has performed best with nearly the accuracy of 98.6%, sensitivity of 0.9720, precision of 1 and specificity of 1.

In the research carried by Pinar Yildirim [8] searches the effect of class imbalance when training data by considering the development of neural network classifier for making medical decisions on chronic kidney disease. In the research, a comparative study was performed using some sampling algorithms based on multilayer perceptron with different learning rate values for the prediction of CKD by considering neural networks. This study reveals that the performance of classification algorithms can be improved by using the sampling algorithms. It also reveals that the learning rate is a crucial parameter which significantly effect on multilayer perceptron.

In the research carried by Asif Salekin and John Stankovic [9] they have introduced a novel approach to detect CKD using machine learning techniques. They have evaluated their research on a dataset with 400 patient records which includes 250 CKD detected patients of early stages with 24 attributes. As the classifiers they have used; k-nearest neighbors, random forest, and neural networks to find an applicable solution. Using a wrapper method they have performed feature reduction analysis to find the attributes which detect CKD with high accuracy and a cost analysis to identify cost effective highly accurate CKD detection classifier by considering 5 attributes which include specific gravity, albumin, diabetes mellitus, hypertension and hemoglobin. The results of this study have introduced new factors to be used by classifiers for detecting CKD more accurately.

III. DATASET AND ATTRIBUTES

This research uses a publicly available dataset [10] which is downloaded from the UCI repository**.** This dataset includes 400 patient records with 25 attributes. Out of those 25 attributes we use a reduced dataset with 14 selected attributes to build the predictive model.

TABLE II. ATTRIBUTES AND THE VALUES USED [7]

Attribute	Value Used
Age	Discrete Integer Values
Blood pressure	Discrete Integer Values
Albumin	Nominal Values(0,1,2,3,4,5)
Red blood cells	Nominal Values(Normal, Abnormal)
Pus cell	Nominal Values(Normal, Abnormal)
Pus cells clumps	Nominal Values(Present, Not-Present)
Serum creatinine	Numeric Values
Hemoglobin	Numeric Values
White blood cell count	Discrete Integer Values
Red blood cell count	Numeric Values
Anemia	Nominal Values(Yes, No)
Classification	Nominal Values(CKD, Not CKD)
Appetite	Nominal Values(Good, Poor)
Packed cell volume	Discrete Integer Values

IV. METHODOLOGY

Data mining is one of a leading technology in the industry. Data mining uses various data mining techniques such as machine learning, artificial intelligence (AI) and statistical. In this study machine learning is used as the technique and the **Cross Industry Standard Process for Data Mining (CRISP-DM)** is used as the research methodology. The CRISP-DM methodology provides a structured approach to planning a data mining project. It is a robust and well-proven methodology. In CRISP-DM there are six major phases [11].

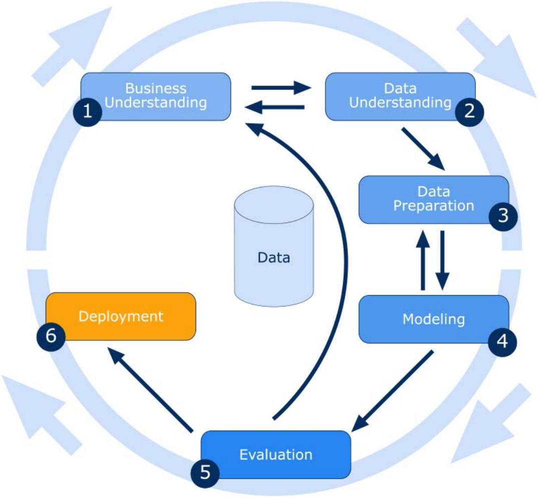

Figure 1. CRISP-DM phases and their processes.

A. Business Understanding

In the business understanding phase, first, it is required to understand business objectives clearly and find out what are the business's needs [11].

In this research project objective is predicting the CKD status (CKD or Not CKD) of the patients to provide early and better treatments for the patients by identifying the CKD status early. For that, we mainly focuses on 14 CKD related attributes such as Serum creatinine level, Hemoglobin level, packed cell volume and etc.

B. Data Understanding

The data understanding phase starts with initial data collection. The data set used in this research was collected from UCI machine learning repository which is a real data set which collected from patients in Tamilnadu, India. This data contains records of 400 instances with 25 attributes to classify and predict chronic kidney disease in early stages.

- Number of Instances: 400 (250 CKD, 150 NOTCKD)
- Number of Attributes: 24 + class = 25 (11 numeric, 14 nominal)
- Missing Attribute Values: Yes
- Class Distribution: ckd and notckd

C. Data Preparation

The data preparation typically consumes about 90% of the time of the project. The outcome of the data preparation phase is the final data set. Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form. The data exploration task at a greater depth may be carried during this phase to notice the patterns based on business understanding [11]. Final data set is sent through the data mining tool to analyze and identify the patterns in the data set. Since this is a real data set this data set contains several missing values. The datasets with missing values are not produce efficient accuracy levels in data mining. To clean the missing values, the missing values of the data set were replaced by zero (0), the default value.

D. Modeling

Under this step different models with different techniques are created with various changers in parameters. The models are created as a solution for an identified problem. The identified problem in this research was not having a proper method to predict the CKD status of the patients using proper attributes. The prediction has been conducted targeting the CKD status (CKD or Not CKD) of the patients. Since this is a classification process the classification algorithms: Multiclass Decision Forest, Multiclass Decision Jungle, Multiclass Logistic Regression and Multiclass Neural Network have been used.

E. Evaluation

In the evaluation phase, the data set is analyzed to identify patterns among the parameters which results the prediction value. At the end of this phase by analyzing and comparing the resulted accuracies of each algorithm, the decision on which algorithm is most suitable to get a better prediction result is achieved. The algorithm with the highest accuracy result was taken as the best algorithm to use with the predictive model. In data mining researches tend to choose the models as the best solutions for the identified research problems with the overall accuracy level greater than 70%. In this study from original data set, filtered data set with 14 attibutes are used in the data mining model. For the testing and training purpose the data set with 400 records was integrated with duplicated data set (copy of original set). Totally there are 800 records. From the whole set, randomly selected 70% of records used to train the model and 30% used for test the trained model.

F. Deployment

The model was developed using Microsoft Azure Machine Learning Studio. The knowledge gained will need to be organized and presented in a way that is useful to the patients and health officers and nephrologists for further decisions based on CKD status of the patients.

V. CLASSIFICATION TASK AND ALGORITHMS

A. *Multiclass Decision Forests*

The decision forest algorithm is an ensemble learning method for classification. The algorithm works by building multiple decision trees and then *voting* on the most popular output class. Voting is a form of aggregation, in which each tree in a classification decision forest outputs a non-normalized frequency histogram of labels. The aggregation process sums these histograms and normalizes the result to get the "probabilities" for each label. The trees that have high prediction confidence will have a greater weight in the final decision [12].

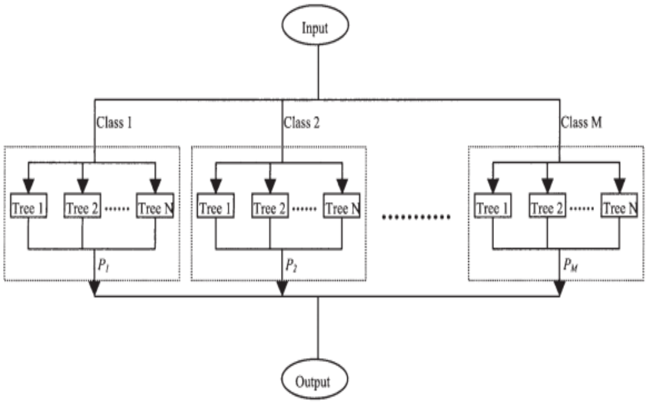

Figure 2. Multiclass decision forest model.

B. *Multiclass Decision Jungle*

The Multiclass Decision Jungle is a machine learning model that is based on a supervised learning algorithm called *decision jungles*. The model can be used to predict a target that has multiple values. Decision jungles are a recent extension to decision forests. A decision jungle consists of an ensemble of decision directed acyclic graphs (DAGs) [12]. A decision jungle J = (G1,...,Gm) is an ensemble of m random decision DAGs G1,...,Gm [13].

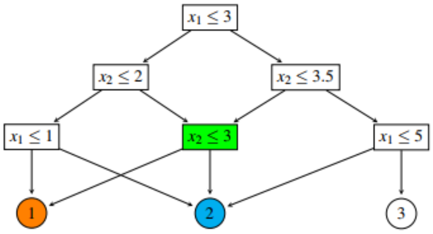

Figure 3. A decision DAG

C. *Multiclass Logistic Regression*

The Multiclass Logistic Regression module to create a logistic regression model that can be used to predict multiple values. Classification using logistic regression is a supervised learning method, and therefore requires a labeled dataset. The label column can contain multiple values. Logistic regression is a well-known method in statistics that is used to predict the probability of an outcome, and is particularly popular for classification tasks. The algorithm predicts the probability of occurrence of an event by fitting data to a logistic function. In multiclass logistic regression, the classifier can be used to predict multiple outcomes [12].

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)} \qquad (1)$$

$$P(Y = 0|X) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i X_i)} \qquad (2)$$

Here $P(X|Y)$ represents the distribution between example X and boolean label Y. In the equation it shows how the logistic regression classifies boolean class label Y [5].

D. *Multiclass Neural Network*

The Multiclass Neural Network module to create a neural network model that can be used to predict a target that has multiple values. A neural network is a set of interconnected layers, in which the inputs lead to outputs by a series of weighted edges and nodes. The weights on the edges are learned when training the neural network on the input data. The direction of the graph proceeds from the inputs through the hidden layer, with all nodes of the graph connected by the weighted edges to nodes in the next layer. To compute the output of the network for any given input, a value is calculated for each node in the hidden layers and in the output layer. For each node, the value is set by calculating the weighted sum of the values of the nodes in the previous layer and applying an activation function to that weighted sum [12].

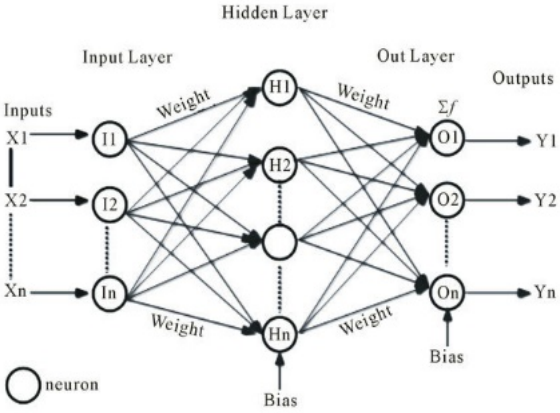

Figure 4. Multiclass neural network model.

VI. EVALUATION AND RESULTS

All the four classifiers were applied to the same dataset using Microsoft Azure Machine learning Studio and the results were obtained and analyzed in the term of predictive accuracy. Predictive accuracy of Z% shows that the classifier is able to classify nearly Z% of instances correctly [7].

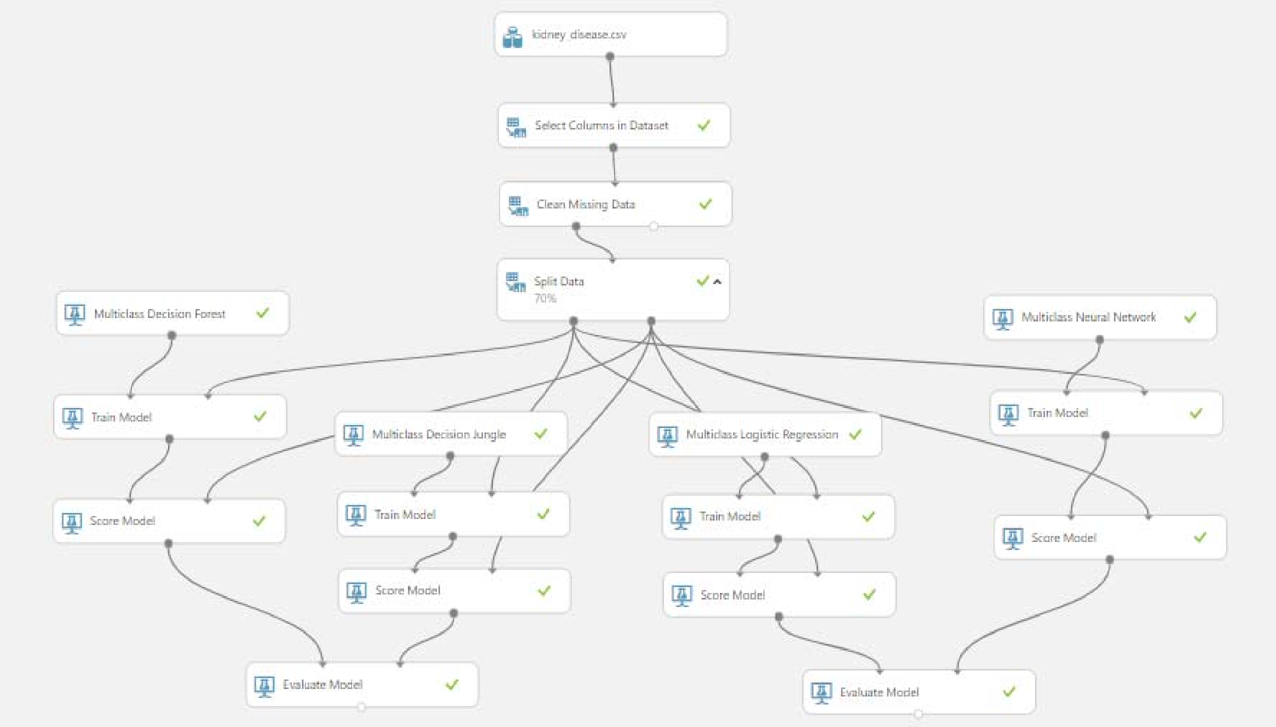

Figure 5. Predictive model created for CKD disease classification.

According to the results obtained using Microsoft Azure Machine Learning Studio it shows that decision forest performs with the predictive accuracy of 0.991which is 99.1%, decision jungle performs with the predictive accuracy of 0.966 which is 96.6%, logistic regression performs with the predictive accuracy of 0.95which is 95.0% and neural network performs with the predictive accuracy of 0.975which is 97.5%.

TABLE III. PREDICTIVE ACCURACIES OF CLASSIFICATION ALGORITHMS ON DATA SET

Used Algorithms	Overall accuracy
Multiclass Decision Forest	99.1%
Multiclass Decision Jungle	96.6%
Multiclass Logistic Regression	95.0%
Multiclass Neural Network	97.5%

From the experimental results obtained, it can be concluded that the decision forest algorithm is appropriated for predicting the CKD status.

VII. CONCLUSION

14 attributes related to CKD patients has been analyzed and predicted for different machine learning classification algorithms: Multiclass Decision Forest, Multiclass Decision Jungle, Multiclass Logistic Regression and Multiclass Neural Network.

From the generated results, it is observed that the Multiclass Decision Forest algorithm provides the highest accuracy of 99.1%. According to the analysis on results, using the above 14 attributes of CKD patients it can predict the CKD status of a new patients with a 99.1% accuracy using this model. The main focus of implemented model is on identifying a new CKD patient's health condition by focusing more featured areas which will help to have a better idea about patient's condition.

The predictive data model is implemented by using different data mining techniques by paying attention to most unpopular data mining algorithms. As per to the literature surveys conducts in this study, it clearly represents that the most researchers use popular data mining algorithms like Naïve Bayes, Support Vector Machine (SVM) and K-Nearest Neighbors as the classification techniques.

According to the earlier researches, most of them have used 24 CKD related attributes to build a predictive model. As an example in research [5] it has obtained SVM as the best classifier to predict CKD with an accuracy of 98.3% by using 24 attributes from the dataset. But according to the results obtained in this research by using only 14 attributes of the same dataset it can achieve 99.1% accuracy using the above created model. Than using 24 attributes, by using less number of attributes with multiclass decision forest algorithm it can achieve a better prediction result with a higher accuracy. In research [7] it has obtained a 98.6% of accuracy by using decision tree algorithm with 24 attributes for the same dataset. When considering the decision tree algorithm it builds the tree based on the entire dataset by using all the features of the dataset but, the decision forest builds using multiple decision trees by categorizing the features of the dataset in to separate trees. The tree with the highest probability have the greater effect on the final result in decision forest. Because of that, decision forest gives more accurate prediction result than the decision tree. Rather than using the decision tree, by using the above model created using the decision forest algorithm it can achieve a 99.1% of better accuracy with less number of attributes.

As an advantage of that, the prediction process is less time consuming. It will help the doctors to start the treatments early for the CKD patients and also it will help to diagnose more patients within a less time period. Limitations of this study are the strength of the data is not higher because of the size of the data set and the missing attribute values. To build a data mining model targeting chronic kidney disease with overall accuracy = 99.99%, will need thousands and thousands of records with zero missing values. But at this level of the study, it will demonstrate the strength of the models by observing statistics metric values which are acceptable in data mining.

VIII. FUTURE WORK

This study can be used as a prototype to develop a healthcare system for CKD patients. To further assess the performance of the model, testing the model with large number of data will help to analyze the accuracy levels of the current model with more accuracy.

ACKNOWLEDGEMENT

We would gratefully acknowledge parents, lecturers of Sri Lanka Institute of Information Technology and the others who provide their valuable guidance and time.

REFERENCES

[1] UFS Health Online, “What is Health Informatics?,” *ufshealthonline.com,* January, 2017.[Online].Available: https://www.usfhealthonline.com/resources/key-concepts/what-is-health-informatics/. [Accessed: Jul. 12, 2017].

[2] Medium, “What’s the relationship between machine learning and data mining?,” *medium.com,* January,2016.[Online].Available: https://medium.com/@xamat/what-s-the-relationship-between-machine-learning-and-data-mining-8c8675966615/. [Accessed: Jul. 12, 2017].

[3] Western Hospital, “Essential Guide to Kidney Disease,” *westernhospital.lk,*2015.[Online].Available: http://www.westernhospital.lk/essential-guide-to-kidney-disease. [Accessed: Feb. 10, 2017].

[4] Naganna Chetty, Kunwar Singh Vaisla, Sithu D Sudarsan, "Role of Attributes Selection in Classification of Chronic Kidney Disease Patients," Proc. IEEE International Conference on Computing, Communication and Security (ICCCS), IEEE, Dec. 2015, doi:10.1109/CCCS.2015.7374193.

[5] Anusorn Charleonnan, Thipwan Fufaung, Tippawan Niyomwong, Wandee Chokchueypattanakit, Sathit Suwannawach, Nitat Ninchawee, “Predictive Analytics for Chronic Kidney Disease Using Machine Learning Techniques,” Proc. Management and Innovation Technology International Conference (MITiCON-2016) , IEEE, Oct. 2016, doi:10.1109/MITICON.2016.8025242.

[6] S.Dilli Arasu, Dr. R. Thirumalaiselvi, "A NOVEL IMPUTATION METHOD FOR EFFECTIVE PREDICTION OF CORONARY KIDNEY DISEASE," Proc. 2nd IEEE International Conference on Computing and Communications Technologies (ICCCT), IEEE, Feb. 2017, doi: 10.1109/ICCCT2.2017.7972256.

[7] Sahil Sharma, Vinod Sharma, Atul Sharma, “Performance Based Evaluation of Various Machine Learning Classification Techniques for Chronic Kidney Disease Diagnosis,” July18, 2016.

[8] Pinar Yildirim, "Chronic Kidney Disease Prediction on Imbalanced Data by Multilayer Perceptron: Chronic Kidney Disease Prediction," Proc. 41st IEEE International Conference on *Computer Software and Applications* (*COMPSAC*), IEEE, Jul. 2017, doi: 10.1109/COMPSAC.2017.84.

[9] Asif Salekin, John Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," Proc. IEEE International Conference on Healthcare Informatics (ICHI), IEEE, Oct. 2016, doi:10.1109/ICHI.2016.36.

[10] L. Rubini, “Early stage of chronic kidney disease UCI machine learning repository,” 2015. [Online].Available: http://archive.ics.uci.edu/ml/datasets/Chronic Kidney Disease.

[11] Smart Vision Europe, “CRISP-DM Methodology”,*sveurope.com,*2016.[Online].Available:http://www.sv-europe.com/crisp-dm-methodology/. [Accessed: May 20, 2017].

[12] Microsoft Azure, “Multiclass Decision Forest,”Microsoft.[Online].Available: https://msdn.microsoft.com/en-us/library/azure/dn906015.aspx. [Accessed: Jul. 26, 2017].

[13] Tobias Pohlen, “Decision Jungles,” July 2014.

[14] Huixiao Hong, Weida Tong, Roger Perkins, Hong Fang, Qian Xie, Leming Shi, “Multiclass Decision Forest—A Novel Pattern Recognition Method for Multiclass Classification in Microarray Data Analysis,” DNA and Cell Biology, vol. 23, no. 10, pp. 685-694, 2004.

[15] Python 3 Codes, “A Neural Network in Python, Part 1: sigmoid function, gradient descent & backpropergation,”Python 3 Codes.[Online].Available: http://python3.codes/neural-network-python-part-1-sigmoid-function-gradient-descent-backpropagation/. [Accessed: Sept. 18, 2017].